A Local Evaluation of the Reliability, Validity, and Procedural Adequacy of the Teacher Performance Assessment Exam for Teaching Credential Candidates

Matthew L. Riggs, Michael P. Verdi, & Patricia K. Arlin
California State University, San Bernardino

In 1998, the California legislature passed Senate Bill SB 2042 which replaced the previous teacher certification legislation. The new law required that all elementary and secondary teacher certification programs in the State of California align their programs with new teaching standards. These standards included Teacher Performance Expectations (TPEs), thirteen performance dimensions defined as relevant to competent teaching (Commission on Teaching Credentialing [CTC], 2008). Every program must also implement a performance assessment to evaluate each teacher candidates' mastery of the TPEs (CTC, 2008). Each teacher certification candidate must pass their program's assessment in order to receive their preliminary teaching credential. Therefore, it was required that all institutions of higher education in the State of California adopt or develop their own version of a performance assessment to be in compliance with the new state law (CTC, 2008). The purpose of this study was to evaluate a pilot administration of this process, providing quantitative and qualitative evidence of its potential.

In order to comply with the legislation, the College of Education at California State University, San Bernardino (CSUSB) chose to adopt the California Teacher Performance Assessment (CalTPA), an approved

Matthew L. Riggs is a professor of psychology and human development, Michael P. Verdi is a professor of educational counseling and psychology, and Patricia K. Arlin is dean, all with the College of Education at California State University, San Bernardino, San Bernardino, California.

assessment program developed by Educational Testing Service (ETS) for the state. Partial administration of the CalTPA began in the Fall Quarter of 2004 on a pilot basis. As the implementation proceeded, it became apparent that some method of tracking and analyzing the data was needed for compliance issues. In the assessment program developed by the CTC (the CalTPA), four tasks are used to measure 12 of the 13 TPEs (CalTPA, 2009). These four tasks are defined and summarized on the CalTPA website http://www.ctc.ca.gov/educator-prep/TPA.html. The scoring model for this assessment program is defined on the website as follows:

> Assessors use rubric-based scoring. Each task has its own dedicated rubric, which assesses the candidate's performance against the TPEs for that task. Rubric scores range from a high of 4 to a low of 1. A candidate must earn a total score of 12 across the four tasks, with no score lower than 2 on any one task. Program sponsors may choose to set a higher passing score than 12.

This global method of scoring results in an ordinal scale of measurement. This level of measurement does enable assessment of inter-rater reliability, and, indeed, a standard for inter-rater reliability is the only psychometric requirement imposed by the CTC upon programs who adopt the CalTPA (CTC, 2008). Interval level measurement, required for more advanced parametric evaluation of psychometric properties, is not achieved within the global, rank-order ratings adopted by the CTC (Anastasi & Urbina, 1997; Salkind, 2006). For example, evidence of factorial validity (reported in the results below) is impossible to evaluate if there are no item-level, dimension specific scores. In addition to the statistical limitations, in global scoring, actual student performance on the individual TPEs is lost as multiple TPEs are combined in each of the four tasks (CalTPA, 2009). This results in the loss of dimension-specific performance data that could be used for program feedback and improvement. For example, if there is a high rate of failure for task four, this could be associated with deficiencies in TPE 1 through 11 or 13. Other than the subjective, qualitative judgments of the raters, there would be no hard data to guide remedial program interventions.

Consequently, it was decided by the multiple-subject (elementary) and single-subject (secondary) program leadership at CSUSB that the data obtained should yield more than just a single, holistic, ordinal score. A dimension-based scoring approach was developed to supplement the holistic scoring. This provided data relevant to the College of Education's Conceptual Framework (CSUSB, College of Education, 2009) for accreditation, as well as enabling better evaluation of the psy-

chometric properties of the measures. While the scoring procedure was supplemental, the teaching performance dimensions used were those defined by ETS for the CTC (CalTPA, 2009).

Both the multiple subject and single subjects programs created two unit assessment courses to prepare and mentor candidates for three of the four tasks—Designing Instruction, Assessing Learning, and the Culminating Teaching Experience (CalTPA, 2009). For these three CalTPA tasks, scoring is handled as part of the assigned coursework. Faculty members teaching the courses do the scoring of the tasks of another instructor's class, and vice versa. The Subject Specific Pedagogy task is the exception. For this task, faculty members are expected to complete a fair share of the scoring during one quarter of the year as part of their service obligation to the program. This practice has now become policy as the legislation was enforced in 2008. The College of Education undertook the process of evaluating its implementation of the CalTPA and the supplemental scoring planned for program improvement data. The results of the study are presented here.

This study utilized quantitative and qualitative approaches to address the following research questions: (1) Did the assessment data possess adequate psychometric properties (including internal reliability, inter-rater reliability, and multiple types of evidence for validity); and (2) Was scoring being completed in an appropriate and consistent manner?

Method

Subjects

Given the pilot status of this program during the data collection period (2004-2005 and 2005-2006 academic years), adoption of the four assessment tasks and completion of the required scoring forms by the faculty was voluntary. Data were collected from all available and usable forms completed during that time period. This resulted in data from 78 secondary and 192 elementary program candidates for Task 1, 42 secondary program candidates for Task 2, 86 secondary and 161 elementary program candidates for Task 3, and 66 secondary and 133 elementary program candidates for Task 4. Demographics were not part of the data collected, so demographic data specific to the sample is not available. However, general demographic data for the population at the time of the study is reported in Table 1. Candidates in the multiple and single subjects programs at CSUSB are predominantly female. There were high proportions of Latino and Caucasian candidates. Candidate ages range higher than many traditional campuses, reflecting a high number of "re-entry" and older first-time candidates.

Table 1
Demographics of Candidates in the College of Education
Fall 2004 and Fall 2005

Variable	*Fall 2004*	*Fall 2005*
Mean Age	36.9	36.7
Gender %		
Women	72.1	71.2
Men	27.9	28.8
Ethnicity %		
Native American	1.7	1.3
African American	13.0	13.7
Hispanic	21.0	23.5
Asian-Pac.Island.	3.8	3.4
White	49.3	46.5
Unknown	11.1	11.5

Measures

The scoring system of the CalTPA required evaluators to assign only a global score of 1, 2, 3, or 4 for each task. Task 1, Subject Specific Pedagogy, was a task that required teacher candidates to read and evaluate case studies and then suggest solutions to the various problems presented (addressing TPEs 1, 3, 4, 6, 7, and 9). Task 2, Designing Instruction, required teacher candidates to plan a lesson for a class and two focus students (including TPEs 1, 4, 6, 7, 8, 9, and 13). Task 3, Assessing Learning, required teacher candidates to plan, implement, and evaluate an assessment of their choosing for their class and two focus students (assessing TPEs 3, 6, 7, 8, 9, and 13). Finally, the Culminating Teaching Experience required the teacher candidates to plan, implement, and videotape both a lesson and its assessment (including TPEs 1 through 11 and 13). After review of each task response provided by the candidate as per the CalTPA program of assessment (CTC, 2008), each rater was to arrive at the global score following a consideration of a set of performance dimensions included within each task. As described above, the performance dimensions originally were defined following a task analysis conducted for the CTC by ETS. These performance dimensions included: (1) planning for assessment, (2) learning about students, (3) making adaptations, (4) goal setting, (5) reflecting on student learning, (6) classroom environment, (7) planning for instruction, (8) pedagogical skill, and (9) analyzing evidence of student learning. ETS had originally proposed that the scoring method would include the creation of thirteen dimension-specific scale scores. This scoring method was deemed "too complex" by the CTC, and the process was altered to

use "dimension-specific rubrics" that would result in a holistic score (CTC, 2003). Using the CTC method, no dimension-level data would be produced. After considering each performance dimension exhibited in the candidate product, the only score assigned by the rater would be the final global rating (CTC, 2008). Because these scores did not provide dimension-specific information, rater forms were prepared to enable raters to provide numeric ratings (1-4) for each performance dimension reflected within each Record of Evidence (ROE). As the ROE is a secure form, it is not possible to provide samples here.

Procedures

Data were collected during the 2004-2005 and 2005-2006 academic years. As described above, data from all candidates evaluated during the voluntary pilot administration were included in this analysis. The sample did not include all candidates in the program, data were not collected from all four tasks for most participants, and this selection process was not random; consequently, this is a non-probability sample. Other than adding dimension-specific scores, test administration and scoring were completed in a manner consistent with that prescribed for the CalTPA (CTC, 2008). One notable limitation to the data was that elementary candidates did not complete the Designing Instruction Task (Task 2) during the data collection period. Though available data from secondary candidates is reported for Task 2, the generalizability of data for this task is clearly limited. For completed forms that had missing data for only one or two performance dimensions, the missing scores were replaced with the modal rating for that facet of performance.

Quantitative Analyses

The analyses and their interpretation in this paper represent a validation study, not statistical tests of correlational hypotheses. Test scores are evaluated within a context of how they are expected to function in reference to absence of error variance (reliability) and whether or not there is evidence that valid inference can be made upon the basis of the test scores (AERA, APA, & NCME, 1999). As such, the concept of statistical significance has reduced relevance. For instance, statistical significance is not an adequate standard for indices of reliability (Anastasi & Urbina, 1997; Thorndike, 2005). In addition, consistent with the concepts of convergent and discriminant validity, the expectation of a null outcome is useful to indicate that the test is working, but difficult to implement within the context of significance testing. Readers unfamiliar with basic validation procedures can find a good introduction to these methods in Salkind, 2006.

Two "layers" of composite scores were developed from the dimension-specific scores. First, the average scores for each conceptual dimension of performance (e.g., planning for instruction, making adaptations, etc.) were calculated. These average dimension scores subsequently were combined to form the Teaching Performance Expectation (TPE) scores. The global score assigned by the rater and a simple mean of all dimension scores also were included in the evaluation. It should be noted that, as per the CalTPA scoring model, the global score is an overall score that represents the most often-used level of evaluation across all dimensions on the exam. Results below include descriptive statistics for all performance dimensions, the TPE's, global scores, and average ratings for all tasks. Given sampling limitations (noted above), not all candidates completed all tasks; nevertheless, mean TPE scores across all tasks were calculated using unweighted averages of the group means. Since no elementary program candidates completed the Designing Instruction Task, Task 2 dimension scores contributed to averages for secondary program candidates, but not for elementary program candidates.

Graphic representations of the mean scores were completed to enable visual comparison between elementary and secondary subject scores, as well as between the first year's (2004-2005) and the second year's (2005-2006) averages. Given that no differences were expected between elementary and secondary program candidates, independent sample t-tests were used to compare secondary and elementary subject scores. Null results would support the validity of the scoring. Finally, psychometric evaluations were conducted on the data, including assessment of reliability (internal consistency and inter-rater), preliminary indicators of validity (factorial validity and criterion validity), rater leniency/severity issues, and convergence between modal (global) and mean scoring approaches.

Qualitative Analyses

Using the qualitative data analysis as described in Miles and Huberman (1994), 100 records of evidence from the Fall Quarter of 2005 were analyzed. Categories were created, each record of evidence (ROE) was read, and evidence was taken from each. To clarify, an ROE is the score sheet that the evaluators use in scoring the CalTPA (CTC, 2008). Again, the only modification to the CSUSB ROE is the addition of dimension-specific scoring. As noted above, the ROE form used for this data was appended to accommodate this change. The main categories for analysis, depending on the task, included the same performance dimensions included in the quantitative scoring and analysis. Next, using this document analysis along with descriptive statistics and personal accounts, the results were triangulated and conclusions were drawn.

Results

Quantitative Results

Descriptives. The range of all mean scores from the Subject Specific Pedagogy Task for both groups (secondary and elementary program candidates) was restricted (2.93 to 3.13 for secondary and 3.00 to 3.19 for elementary subjects (see Table 2). There appeared to be no meaningful differences in magnitude among the performance composites. None of the between groups t-tests comparing secondary to elementary program candidates were significant. Comparison of all data from both years did not suggest meaningful change from year-to-year.

Greater variability existed among the dimension and TPE scores from secondary program candidates on Designing Instruction (ranging from 2.72 to 3.25, see Table 3). "Making adaptations" appeared to have been the lowest score from all samples. The second-year secondary program candidates scored peak performances on "Learning about Students," "Reflection," and "TPE 8" (which is based exclusively on the "Learning about Students" performance dimension). The "Learning about Students" score for the second year's secondary program candidates appeared to be a marked improvement over the previous year's average.

For Task 3, Assessing Learning, some variance existed within groups across dimensions (means ranging from 3.37 to 3.76 for secondary program candidates and from 2.75 to 3.26 for elementary program candidates (see Table 4), but the greater source of variance appeared between the two groups. Inconsistent with expectations, independent t-tests on all

Table 2
Task 1 (Subject Specific Pedagogy) Descriptive Statistics

	Elementary Program Candidates		Secondary Program Candidates	
Score	M	SD	M	SD
Global Score	3.15	0.62	3.08	0.73
Mean Score	3.09	0.49	3.03	0.57
Planning for Instruction	3.19	0.50	3.13	0.51
Planning for Assessment	3.04	0.53	2.99	0.63
Making Adaptations	3.01	0.59	2.93	0.70
Pedagogical Skill	3.00	0.67	2.93	0.76
TPE 1	3.07	0.53	3.00	0.62
TPE 3	3.04	0.53	2.99	0.63
TPE 4	3.10	0.50	3.03	0.58
TPE 6	3.07	0.53	3.00	0.62
TPE 7	3.10	0.50	3.03	0.58

Table 3
Task 2 (Designing Instruction) Descriptive Statistics

	Secondary Program Candidates	
Score	M	S
Global Score	3.17	0.76
Mean Score	3.01	0.62
Goal Setting	2.90	0.65
Learning about Students	3.25	0.65
Planning for Instruction	3.02	0.86
Making Adaptations	2.72	0.79
Reflection	2.88	0.87
Pedagogical Skill	3.14	0.74
TPE 1	3.01	0.63
TPE 4	2.97	0.62
TPE 6	2.94	0.71
TPE 7	2.97	0.66
TPE 8	3.25	0.65
TPE 9	2.88	0.65
TPE 13	3.01	0.75

dimensions and TPE averages showed statistically significant differences between elementary and secondary program candidates. While a difference was apparent between the two groups during 2004-2005,

Table 4
Task 3 (Assessing Learning) Descriptive Statistics

	Elementary Program Candidates		Secondary Program Candidates	
Score	M	SD	M	SD
Global Score	3.12	0.61	3.65	0.55
Mean Score	3.01	0.53	3.51	0.42
Goal Setting	3.26	0.66	3.76	0.53
Planning for Assessment	3.16	0.58	3.47	0.58
Learning about Students	3.07	0.71	3.55	0.54
Making Adaptations	2.75	0.76	3.37	0.53
Anal. Evid. of Learning	3.01	0.60	3.58	0.52
Reflection	3.05	0.65	3.53	0.51
TPE 3	3.07	0.50	3.53	0.46
TPE 6	3.01	0.53	3.57	0.41
TPE 7	2.91	0.66	3.46	0.49
TPE 8	3.07	0.71	3.55	0.54
TPE 9	3.11	0.52	3.63	0.44
TPE 13	3.05	0.65	3.53	0.51

this performance/scoring gap increased for 2005-2006. Within groups, both elementary and secondary program candidates scored highest on "Goal Setting" and lowest on "Making Adaptations."

For the Culminating Teaching Experience (Task 4) scores, secondary program candidate scores were higher than elementary program scores in all categories (based on independent sample t-tests). Means are reported in Table 5. Secondary program averages ranged from 3.46 ("Making Adaptations") to 3.80 ("Goal Setting), while elementary program scores ranged from 2.91 ("Making Adaptations") to 3.37 ("Planning for Instruction"). These discrepancies between group scores did not exist in 2004-2005 groups. While the difference in sampling procedures between the two years has been noted, the reason for such a change would not seem obviously related to any systematic sampling bias. It also must be noted that this difference in scores between the two groups was not apparent in the evaluations from Subject Specific Pedagogy.

Table 5
Task 4 (Culminating Teaching Experience) Descriptive Statistics

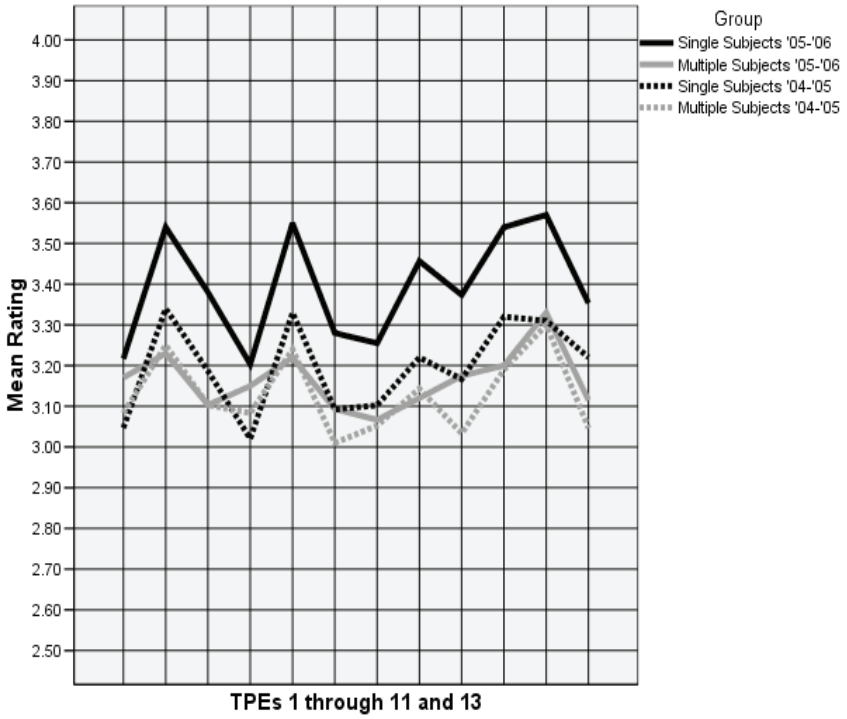
	Elementary Program Candidates		Secondary Program Candidates	
Score	M	SD	M	SD
Global Score	3.31	0.55	3.67	0.54
Mean Score	3.17	0.52	3.56	0.41
Goal Setting	3.23	0.59	3.80	0.36
Learning about Students	3.17	0.63	3.57	0.43
Classroom Environment	3.33	0.63	3.57	0.46
Planning for Instruction	3.37	0.60	3.59	0.49
Making Adaptations	2.91	0.74	3.46	0.56
Pedagogical Skill	3.32	0.57	3.58	0.61
Anal. Evid. of Learning	3.15	0.66	3.48	0.54
Reflection	3.21	0.64	3.55	0.61
TPE 1	3.28	0.48	3.63	0.41
TPE 2	3.23	0.55	3.54	0.45
TPE 3	3.20	0.54	3.61	0.42
TPE 4	3.20	0.52	3.60	0.39
TPE 5	3.22	0.50	3.55	0.42
TPE 6	3.21	0.50	3.60	0.42
TPE 7	3.19	0.52	3.55	0.43
TPE 8	3.17	0.63	3.57	0.43
TPE 9	3.24	0.52	3.61	0.42
TPE 10	3.20	0.55	3.54	0.43
TPE 11	3.33	0.63	3.57	0.46
TPE 13	3.18	0.60	3.52	0.51

Total TPE scores as averaged across all four tasks are illustrated by group in Figure 1. Variability within groups during the 2005-2006 year was about the same for both groups, with performance profiles running nearly parallel with each other. Again, the variance between groups is apparent for the second year, but not the first year.

Evidence of Reliability

Reliability was first evaluated in reference to internal consistency. Cronbach alphas were calculated on each set of items by task. The initial solution attempted was based upon entry of all items across all dimensions as though they were measuring a single construct. Alphas based upon the "single measure" approach were very high (.95 for Subject Specific Pedagogy, .95 for Designing Instruction, .95 for Assessing Learning, and .96 for the Culminating Teaching Experience). These values indicate that all items can be used as a measure of a single, global construct. While it is convenient that all items can be used together as a single measure,

Figure 1
Total TPE Results from 2004-2005 and 2005-2006



one might have expected greater discrimination in ratings among the different performance dimensions. This will be further investigated as part of the factorial validity analysis to follow in the next section.

Inter-rater reliability also was calculated, and the results evaluated. This was achieved by having the College of Education TPA Coordinator randomly select five tests graded by each scorer and re-grade them in a blind control. This resulted in 80 tests being re-scored by the TPA Coordinator. Upon completion of the re-grading portion of the analysis, the new ROEs were correlated with the original ROEs. Intra-class correlations (ICCs) and Pearson correlations were computed between the Director's numerical ratings and those obtained from the original scorer. ICC's are more rigorous coefficients of reliability in that they can analyze both covariance (whether the scores go up and down together), and absolute agreement (whether ratings are at the same level) (Shrout & Fleiss, 1979; Cicchetti, 1994). The Pearson correlations indicate only whether scores co-vary (Howell, 2010). The results of these analyses are reported in Table 6. ICC's are interpreted differently than estimates of internal consistency such as Cronbach's alpha (Cicchetti, 1994). Cicchetti's standards will be employed here: below .40 is poor, .40 to .59 is fair, .60

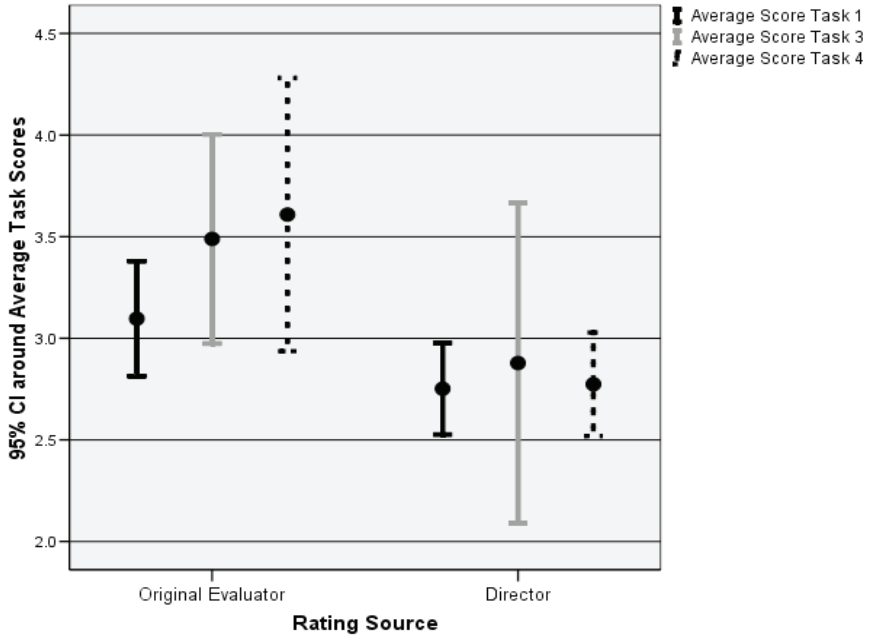
Table 6
Inter-rater Reliability Coefficient

Score	ICC	Pearson's r
Global Score Task 1	0.25	0.41
Mean Score Task 1	0.28	0.66
Global Score Task 3	0.41	0.61
Mean Score Task 3	0.23	0.43
Goal Setting Task 3	0.18	0.35
Planning For Assessment Task 3	0.04	0.07
Learning About Students Task 3	0.27	0.44
Making Adaptations Task 3	0.21	0.40
Analyzing Student Evidence & Assessment Task 3	0.37	0.48
Reflection Task 3	0.32	0.38
Global Score Task 4	0.27	0.36
Mean Score Task 4	0.32	0.42
Goal Setting Task 4	0.08	0.13
Learning about Students Task 4	0.31	0.37
Classroom Environment Task 4	0.10	0.10
Planning for Instruction Task 4	0.41	0.47
Making Adaptations Task 4	0.19	0.24
Pedagogical Skill Task 4	0.01	0.02
Analyzing Evidence of Student Learning Task 4	0.33	0.39
Reflection Task 4	0.45	0.51

to .74 is good, and .75 and above is excellent. Given the single factor solution for the Subject Specific Pedagogy Task (see results in validity section below), only the global and overall mean scores for this task were tested. Designing Instruction was not included due to the limited sample. The coefficients were calculated for global and mean scores, as well as all performance facets for both the Assessing Learning Task and the Culminating Teaching Experience Task.

While some of the correlations are statistically significant, tests of statistical significance are not applied to reliability coefficients (Anastasi & Urbina, 1997; Thorndike, 2005). The results for the ICC's in Table 6 generally indicate *in*adequate inter-rater reliability. This conclusion does not necessitate the contention that one rater is correct and the other is wrong. It simply provides evidence that the two sets of scores do not show adequate agreement, and that very different scores might be awarded to the same student product when evaluated by different raters. Also, a systematic difference in level did exist between the director's ratings and the average ratings of the evaluators, with the director's ratings running below the means of the actual raters ($t = 2.65$, $p = .029$

Figure 2
95% CIs around Average Ratings Obtained for ICCs



for Subject Specific Pedagogy; $\underline{t} = 4.37$, $\underline{p} < .001$ for Assessing Learning; and $\underline{t} = 3.31$, $\underline{p} = .001$ for the Culminating Teaching Experience). These differences are illustrated in Figure 2.

The Pearson r's also are much lower than might be expected, indicating that the problem is not limited simply to a difference in level (leniency/severity), but also to a lack of simple covariance between scores (agreement as to which students were best and which were worst). A question was raised, however, during subsequent meetings regarding these results. Do these low coefficients of reliability indicate disagreement among all raters, or potentially only some? We did not conduct a full retrospective analysis of the agreement between the director's ratings and each of the evaluators individually, but differences in agreement between the director and trained raters were graphically compared with those of untrained or only partially trained evaluators. Figures 3 and 4 illustrate a rather common observation. Figure 3 shows the ratings of the Director for the Assessing Learning Task compared with a trained evaluator. Figure 4 illustrates the Director's scores with those of an

Figure 3
Directors' Ratings with Ratings of a Trained Evaluator

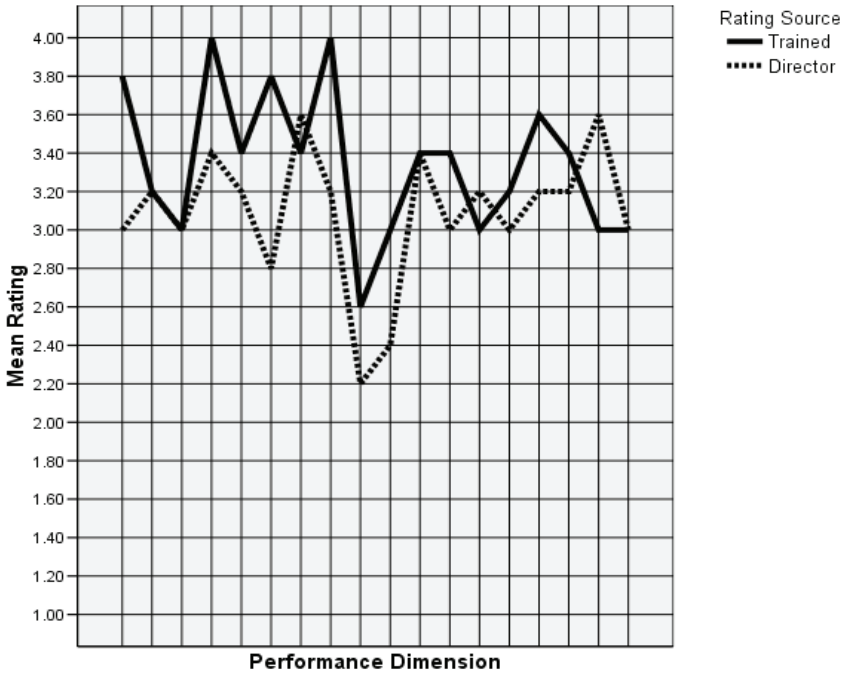
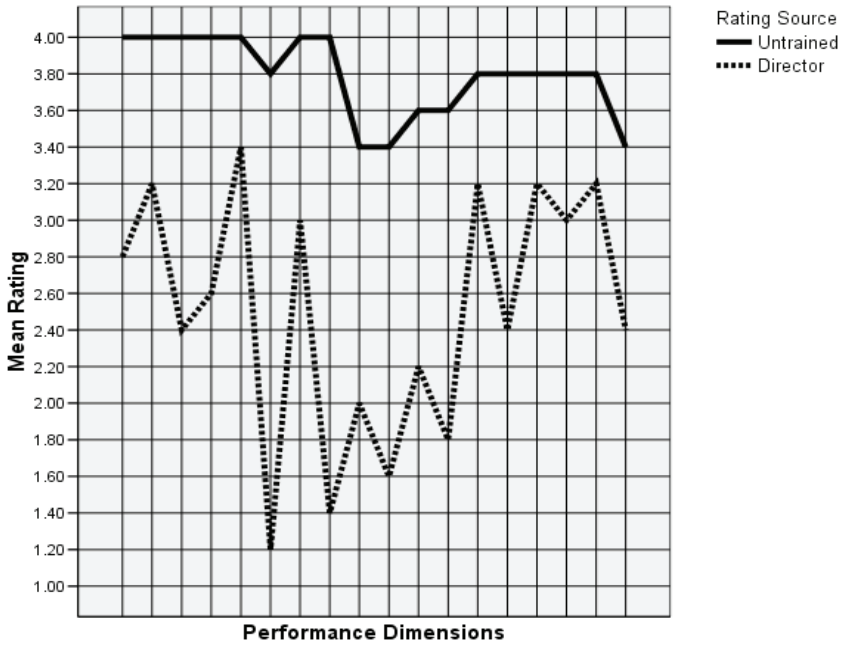


Figure 4
Director's Ratings with Ratings of an Untrained Rater



untrained evaluator for the same task. Though only partial evidence, these would appear to indicate that inter-rater reliability is not only possible, but perhaps already adequate among those properly trained. At this pilot point of implementation, evaluators had not all been fully trained, or had attended training but never completed "calibration" (critical to adequate training).

Evidence of Validity

The first analysis used to provide evidence of validity was factor analysis of the items. If items intended to measure common performance dimensions form factors as intended, this provides evidence that items are measuring distinct constructs, as well as measuring the number of constructs expected. Though a pre-defined theoretical structure of the number of factors existed, because structures were previously untested, exploratory factor analysis was attempted first. Preliminary attempts to factor analyze the items scored for the Subject Specific Pedagogy Task indicated a lack of factorial validity. Only a single factor solution could be extracted. This indicated that scores on Subject Specific Pedagogy

may all measure a single construct, and discriminant measures of subfactors (i.e., planning for instruction, making adaptations, etc.) might not be practical.

The Designing Instruction Task was not factor analyzed due to lack of data for this sample. For the Assessing Learning Task and the Culminating Teaching Experience Task, a principal axis factor extraction with a promax rotation was used, and the expected number of factors was extracted. The resulting solutions approximated the intended pattern of factors very closely. The only deviation from the intended set of factors for both solutions was that the "Learning about Students" items formed two factors. For the Assessing Learning Task, the single "Goal Setting" item did not form a coherent factor of its own. The same happened with the single "Pedagogical Skill" item in the Culminating Teaching Experience Task. This result is not unusual for a single item, and could be fixed easily by adding one or two additional items to measure these dimensions.

Because there was a pre-defined structure of measurement, a confirmatory factor analysis (CFA) also was applied to the same data from the Assessing Learning Task and the Culminating Teaching Experience Task. These results also provide strong evidence that, at least for these two tasks, six and eight distinct constructs are being measured respectively, and items are functioning together as intended. For both CFA's, the Chi-square, though significant, met the 2:1 ratio with the degrees of freedom, and the comparative fit index (CFI)'s of .97 (Assessing Learning) and .96 (Culminating Teaching Experience), as well as the root mean square error approximation (RMSEA)'s of .06 (both these tasks) indicated a good fit of the data to the measurement model.

Validity was further investigated by correlating student performance on the TPE's with Grade Point Averages (GPA). While GPAs do not provide an ideal variable for convergent validity, one might expect these two constructs to correlate at some level. This expectation was not, however, borne out in the resulting analysis (correlations shown in Table 7). With the exception of Designing Instruction (which represents a limited subset of single subject candidates only), none of the correlations approach a magnitude that might indicate a meaningful relation between GPA and performance on the TPA's. While correlations with the Designing Instruction Task were moderate to large in magnitude, those for the Subject Specific Pedagogy and Assessing Learning Tasks were generally small, and those for the Culminating Teaching Experience Task were very close to zero. This may not be cause for alarm, as the expectation that GPA would converge with TPA performance may be wrong or simply an artifact of a restricted range of GPAs for fifth

year students. In fact, if TPA's were intended to measure constructs not dependent on general academic skill, this result would support the discriminant validity of these scores. Eventually, better outcomes, such as evaluations of actual teacher performance in the classroom should be used for the purpose of predictive criterion validation.

Rater Severity/Leniency Issues

To investigate the degree to which different raters appeared to be applying approximately equal standards, graphic displays of the central tendency and dispersion of ratings by task by rater were created. Figures reflecting these data for the Assessing Learning Task and the Culminating Teaching Experience Task are provided (see Figures 5 and 6, respectively). The assumption is that all raters would have candidates of approximately equal aptitude (not necessarily true, but assumed barring evidence otherwise). Consequently, the psychometric expectation would be that, if raters are applying equitable standards, raters would produce scores with approximately the same means with overlapping dispersion. The graphs, as well as subsequent analysis of variance explained by the source of rating, indicated that this was not always true. Figures 5 and 6 illustrate 95% confidence intervals around the mean rating for each grader. Multiple subject raters are to the left (open circle for mean) and single subject raters are to the right (solid circle for mean).

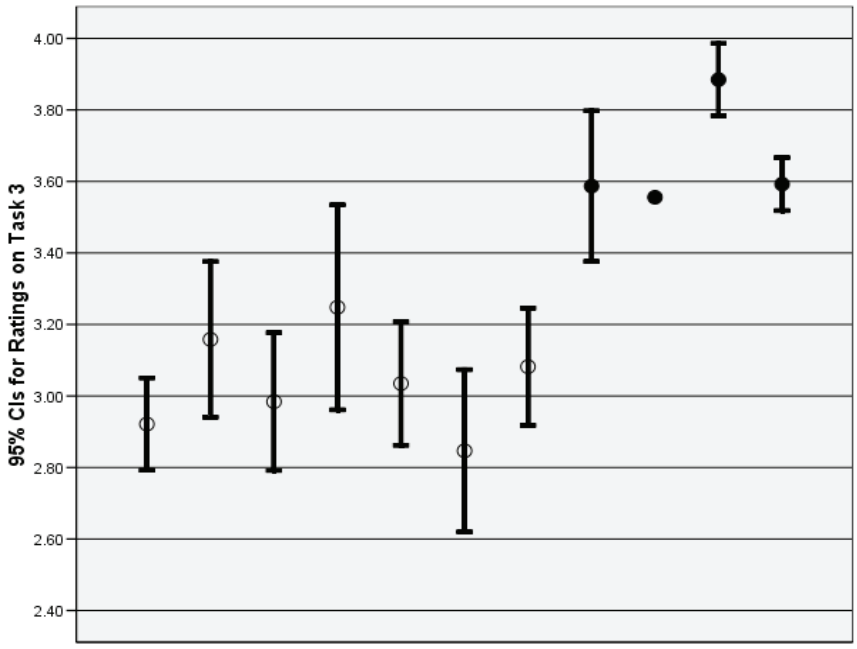
All single subject means are greater than all multiple subject means. The probability of this rank ordering of means by programs happening by chance is .008 for Assessing Learning and .006 for Culminating Teaching Experience. Also, ANOVA models using assessor as the predictor and mean scores as the criterion are significant for both Assessing Learn-

Table 7
Correlations between TPEs and GPAs

Score	GPA Cumulative	GPA Last 60-90
Global Score Task 1	.15*	.17*
Mean Score Task 1	.18*	.19*
Global Score Task 2	.34	.49
Mean Score Task 2	.34	.53*
Global Score Task 3	.15	.19*
Mean Score Task 3	.12	.17
Global Score Task 4	.01	.02
Mean Score Task 4	.03	.05

* $p < .05$

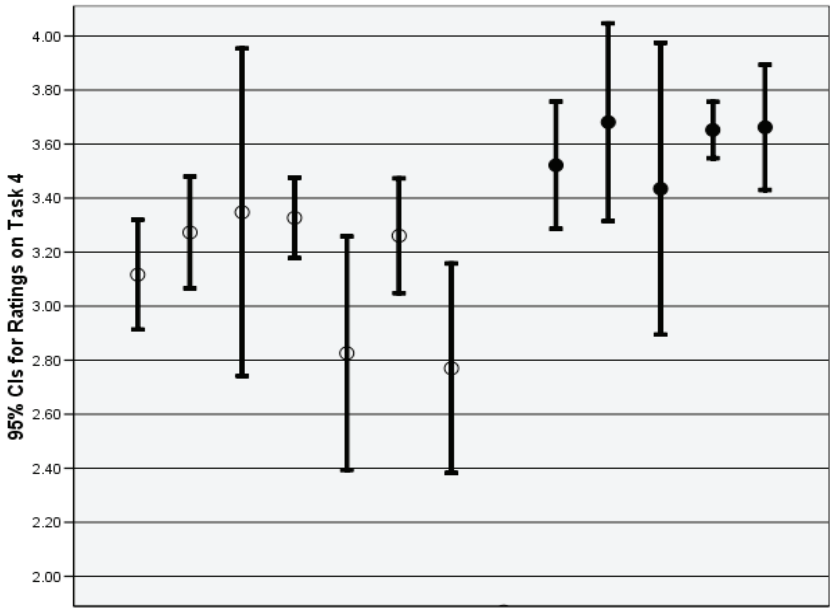
Figure 5
95% CIs for Mean Ratings by Raters by Program for Task 3



ing and Culminating Teaching Experience (F (9,210) = 8.46, p < .001; F (12,162) = 4.07, p < .001 respectively). MANOVA was not applied for two reasons. First, as described in the description of sampling, data was not available for all candidates across all tasks. Listwise deletion would have reduced the sample size by two-thirds. Second, because weighted linear composites based upon correlated scores are prone to multicolinearity, such composites should be avoided as multiple criteria, as they are for multiple predictors in regression analyses (ref.). For Assessing Learning, 27% of the variance is explained by rater, with 23% of the variance explained by rater for the Culminating Teaching Experience Task. Most of this variance can be attributed to program (single vs. multiple) with 19% of the variance in ratings explained by program for Assessing Learning, and 12% of the variance explained by program for the Culminating Teaching Experience.

Assuming no systematic differences in actual quality of candidates by program and/or by rater, if no leniency-severity issues existed, one would expect that no appreciable variance would be explained by program or by rater. Granted, many potential confounds may contribute

Figure 6
95% CIs for Mean Ratings by Raters by Program for Task 4



to these results, but the overall pattern of results, especially given the systematic difference in training quality by program (described in the inter-rater reliability section above) support the interpretation that completion of training (especially the calibration portion of the training) would reduce the magnitude of these effects to negligible, non-significant differences.

Mode vs. Mean Scores

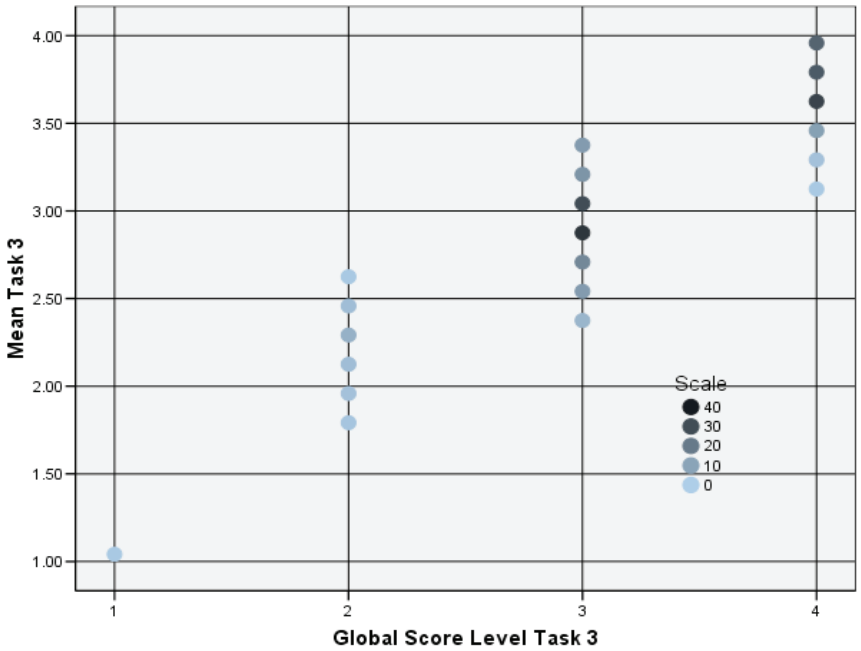
A final issue relevant to the validity of the inferences to be based upon the TPA scores involves a consideration of whether or not pass/fail decisions should be based upon the mode or the mean of scores across performance dimensions. The original CTC scoring model was based upon modal scoring only with a candidate required to score 12 out of 16 possible points on the global scores summed across each of the four tasks. This standard was locally adjusted to require the student to score a mode of "3" on every individual task. However, given that other local adjustments were made to ensure that numerical scores were assigned for each performance item, the calculation of a mean seemed justifiable.

While the state may continue to promote passage based upon global

scores, mean scores, with their superior psychometric properties, often are easier to defend in grievance procedures. Figure 7 used the mode and mean scores from the Assessing Learning Task to illustrate the potential for great dispersion in mean scores for those with identical mode scores. This graph shows that candidates receiving a modal score of 4 actually have mean scores as low as 3.1, while some candidates receiving a model score of 3 actually have mean scores as high as 3.4. A problem that might cause concern specific to the validity of the pass/fail decision is the circumstance illustrated here where a student with a global score of 2 (failing) actually has a mean score closer to 3 than 2 (2.6).

Other illogical, difficult to defend circumstances may occur when modes are used rather than means. If a student has 12 "4's" and 10 "3's" on their first 22 of 25 ratings, they end up receiving a higher modal rating if the remaining three scores were "2's" rather than "3's." All else being equal, when parametric assumptions are approximated (e.g., normality and absence of serious outliers), means are easier to defend because they utilize all numbers in the distribution in a more equitable fashion.

Figure 7
Mean Scores by Modal Scores, Task 3



Qualitative Results

Overall Score

The qualitative analysis revealed four areas of concern in grading and administering the CalTPA. First, the overall or modal score given on 16.5% of all the tests taken were incorrectly scored by the initial assessor. For the approximately 400 exams in the second quarter of the first year of the study, 66 candidates received the incorrect score. More often then not, candidates received a grade lower than they should have based on the scorers' evaluation of each dimension. The average individual scores given were very often averaging close to 4, yet the overall grade was closer to a score of three.

The vast majority of the candidates passed the exam; therefore few complaints were registered. This outcome does raise the question, however, if the graders, the tests, or both were biased in favor of our candidates. It was concluded that grading practices needed to be re-evaluated. This is consistent with conclusions reached on the basis of the quantitative evaluation of inter-rater reliability.

Questionable Grading Practices

The Records of Evidence (ROE) also revealed that some of the assessors were using questionable grading practices and judgments while scoring. First, assessors sometimes did not provide sufficient evidence for the evaluation they made. Occasionally, opinion was substituted for fact on the evidence side of the ROE. In the extreme case, on two ROE's, all 28 of the evidence lines lacked sufficient or correct evidence, yet both papers received a score of 4. It is critical that each line of evidence contains information that can be used by the original teacher so that they can explain the ROE and provide adequate feedback to the student. Assessors must be made aware that they are looking for evidence contained within the original test paper. Opinions should not be inserted. The evaluation side of the ROE is to be used for that purpose.

Second, in eight percent of the ROE's, the original grader failed to evaluate at least one of the performance dimensions. This forced the research team to substitute the modal value for these missing values so that their data could be used. However, these scores may or may not have reflected the actual performance of the student and hence contaminates the overall scores. It is clear that in a modal grading system, a single evaluation can and often does affect the overall grade that a student can receive. Due to the sensitivity of the modal system, missing evaluations cause large problems for both the outcomes assessment, as well as the actual scoring of the task.

Third, in several instances, the ROE's included incorrect answers or questionable evaluations. For example, in twenty-seven percent of the ROE's, evaluators rated answers that contained more than one selection from a list of multiple options when the question clearly stated that they should pick and defend one choice. In some cases, every choice from the list was chosen. In this case evaluators should have rated these answers as 1 or 2, but often did not.

A final questionable grading practice found among the ROE's was the actual changing of grades by the instructor of record. In four percent of the ROE's evaluated, the modal score had been changed from that assigned by the original independent evaluator's score. Apparently, this was due to the fact that the original instructor felt the score assigned by the independent scorer was inadequate or did not reflect student effort. Student awareness and abuse of this practice has had serious consequences. To remedy this problem, results are no longer returned to the classroom instructor prior to them being reviewed by the TPA office. Any task that is scored as a 1 or 2 (failing) automatically receives a second scoring from an assessor assigned specifically to this task. At this time, if the two scorings are not on point or equal, the test in question is scored a third time by the TPA Coordinator. A single score is then assigned based on all three scores. All tests are returned to the TPA office, and the results are forwarded to the classroom instructors so that they can be returned to the teacher candidates. These revised procedures have solved the problem of classroom instructors changing initial scores.

Additional Safeguards and Improvements

All assessors met with the TPA Coordinator and practice exams were scored together to ensure grading procedures were followed and all questions were answered. Additional safeguards added now require the assessor to make sure that evaluations were made for each aspect of teaching, and that the final score placed on the ROE matches the work sheet that the assessor must fill out before completing their scoring of a task. Finally, it was established that each assessor would have a random sample of their scores rescored by the TPA Coordinator each quarter. Assessors must have an 80% agreement of the five scores and 50% of the scores must be on point. These safeguards have resulted in a decline in mismatch and incorrect scores and a dramatic increase in inter-rater reliability. In the most recent evaluation of assessors (Winter Quarter 2007-2008), the assessors achieved 94.8% agreement between the first and second scorer and 78% of all tasks scored by the assessors were on point.

Outcomes Assessment

As they existed in the first years of piloting, ROE data did not enable full qualitative evaluation of the scores' potential to serve the purpose of outcomes assessment. Since the ROEs often were incomplete and did not give a full picture of what our candidates actually were doing, it was determined that the ROE data would not be used for outcome assessment at this time. Instead, it was suggested that a random sample of the actual tasks be used in future qualitative analysis.

A few preliminary findings were made from the examination of the ROE's. First, it is apparent that little original thought was being required on some of the CalTPA tasks. This is particularly true of Subject Specific Pedagogy. In over fifty percent of the Multiple Subjects ROE's reviewed, evaluators noted that candidates would use a Venn Diagram to teach similarities and differences, reading the textbooks as the difficult assignment, and using grade level materials as an adaptation. In fact, four other answers were repeated on at least 50% of all the ROE's. This pattern also was evident for Single Subjects. The exception was only apparent when comparing those ROE's from teacher candidates with the same major. For example, the tests of history majors were similar to those of other history majors but varied from all other subject matter areas. These practices may explain some of the psychometric problems seen with this specific task.

A second trend seen in the ROE's was that several inappropriate adaptations were suggested for Special Needs Students. In fact, Making Adaptations appeared from the document search to be one of the weakest areas on the exams. Quantitative analysis supported this assertion. A final problem found from the examination of the qualitative data was possible plagiarism. It appears that approximately 7-10 % of our sample from the Spring Quarter 2006 contained at least one answer that appeared to be plagiarized. In reference to the Subject Specific Pedagogy Task, fifty percent of the ROE's contained at least seven content-identical answers. The problems with this task are well documented. Additional versions of the Subject Specific Pedagogy Task are currently under development by the state to correct this problem. Since the same limited number of questions and answers are available, similar answers are the logical result.

The remaining three tasks pose a different problem than the Subject Specific Pedagogy Task. Here, evidence of plagiarism comes in two forms. The first is that the gender of the focus student will change from one paragraph to the next. This indicates that candidates are cutting and pasting several answers together. The second form of plagiarism that is found in the remaining three tasks is the reuse of the same answer

to fill each box for the focus students. Candidates were simply pasting the answer for the first student in for the second, and subsequently for the whole class. In these cases, evaluators must be more vigilant when scoring and report the discrepancies they find. Appropriate guidelines and rules need to be written to cover these occurrences so that a fair and balanced approach can be applied.

To that end, all teacher candidates are now required to attend a Teaching Performance Assessment Introductory Seminar held at the beginning of each quarter. At this meeting, candidates are told of the requirements of the teaching performance assessments, and the University policy concerning plagiarism is reviewed. Upon completion of the meeting, candidates are required to sign a form that states that they are aware of the policy and understand that the consequence for breaking the policy can be removal from the teaching credential program and expulsion from the University. These forms are kept in the teacher candidate's TPA file.

Recommendations and Conclusions

In recent years, teacher education has come under fire in reference to the adequacy of teacher training and the levels of accountability to which pre-service and intern teachers are held. Leaders in the reform movement of teacher education have called for programs to be more challenging, intensive and accountable (Darling-Hammond, 1997; Goodlad, 1994). One way that the state of California has responded to the reformers was through the adoption of the CalTPA and the subsequent research that has been conducted on the exam and its implementation. We believe that the results of this study support the following recommendations.

First, it is clear from both a quantitative and qualitative perspective that the use of a global or modal score as the only score a teacher candidate receives limits the potential value of the exam data. Not only are modal scores sensitive to idiosyncratic combinations of performance dimension evaluations, they provide little (perhaps no) potential for meaningful feedback to the candidate or the program. Global scores only provide evidence that candidates are failing, but not why or how. Given our experience, it has been useful to create a scoring procedure that enables raters to document a numerical value for each performance dimension. These dimensions were developed by a highly competent organization (ETS), and our data supports the factorial validity of raters' scores by dimension. The dimension scores also can provide quantitative support of the global rating currently used for the pass/fail decision.

Also in reference to creating ROE's and scoring, our initial results

led to the recommendation that the ROE should be completed via an electronic format on a secure website rather than on paper. Not only has this ensured that all dimensions are scored, and that modal scores are correctly determined, we believe it has reduced the workload associated with completion of the ROE. Keywords from the rubric are readily available. Evidence provided in electronic media from the student can be cut and pasted rather than transcribed. Data is easily reviewed and edited. Since numerical ratings are already in an electronic database, potential for data entry errors are eliminated. This recommendation has been successfully implemented at CSUSB.

Next, it is recommended that the all procedures concerning the CalTPA be standardized and clearly communicated in order to reduce ambiguity and inconsistency. The areas in which standardization are needed include but are not limited to the training of assessors, the presentation of information to the candidates, and in reference to procedures for student appeals of scores they received on the exam. A central policy manual should be written that defines procedures for each of theses areas and distributed to the appropriate audience. Presently, CSUSB has implemented this suggestion with great success.

Perhaps the most critical aspect of need for standardization is in the area of assessor training. ALL assessors are now trained and evaluated periodically through the school year. A random sample of their scores is rescored by a second assessor. An 80% percent global score agreement (with 50% of the dimension scores being on point) must be achieved if they are to continue as an assessor for the program. Moreover, all assessors are now required to meet with both the TPA Coordinator and their respective program directors to review scoring practices and to discuss difficult examples that occurred while they scored. This strengthening of procedures was deemed necessary due to the initial low values obtained for quantitative assessment of the inter-rater reliability.

We also have improved candidate orientation in reference to TPA procedures and requirements. Instructional presentations have been created. These are presented to new and continuing teacher certification candidates at required informational meetings. Moreover, information is given to candidates in the form of official flyers that are distributed in their assessment courses. Also, those teaching the assessment courses are required to review all procedures for taking the test and for selecting appropriate focus students. This discussion is done at the beginning of each quarter. This has eliminated much confusion and anxiety surrounding the test and the requirements candidates must meet. In the area of student appeals, a committee consisting of the TPA Coordinators, the two Program Directors, and the Associate Dean of Teacher Education

have met and developed an appropriate appeals process for the CalTPA at CSUSB. These procedures will be placed into the student handbook of each program, and they are presented to candidates during the Subject Specific Pedagogy meeting that is held each quarter.

Finally, it is recommended that the research into the TPA process should continue, both qualitatively and quantitatively, to provide data on the reliability and validity of the process, the meaning of the scores, and the adequacy of inferences based upon those scores. Furthermore, we believe that all who choose to use this battery of exams as their teacher performance assessment should do their own research and contribute data to support (or refute) the reliability and validity of this process. The process of determining construct validity is never finished (Shultz, Riggs, & Kottke, 1999; AERA, APA, & NCME, 1999), but accumulation of data over multiple sites across multiple years will enable improvement and refinement of assessment, and contribute to the accountability that society expects of teacher preparation programs. Future research also must include specific criterion validation tests using true teaching performance outcomes. This institution currently is investigating the predictive validity of these assessment scores for evaluations of credential candidates' performance during student teaching. Research eventually should evaluate the predictive validity of the scores for teaching performance of first year teachers.

In conclusion, the research presented here gives an initial look at the process of a college of education piloting and assessing a high stakes exam that candidates must now take and pass in order to become a certified and credentialed teacher in the state of California. With the passing of SB 2042 and SB 1209 (which more recently made implementation of a teaching performance assessment mandatory starting July 2008), we hope that colleges and universities that train teachers can use the information presented here to improve their own implementation and administration of the exam.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

California Senate Bill SB 2042 (1998). Website materials downloaded 2/22/2009 from http://www.leginfo.ca.gov/pub/97-98/bill/sen/sb_20012050/sb_2042_bill_19980918_chaptered.html

California State University San Bernardino. College of Education. (2009). *College of Education's conceptual framework.*

California Teaching Performance Assessment. (2009). Website materials downloaded 3/10/2009 from http://www.ctc.ca.gov/educator-prep/TPA-files/CalTPA-general-info.pdf.

Commission on Teacher Credentialing (CCTC). (2008) *California teaching performance assessment foundations day: Assessor training handbook.* Sacramento, CA: CTC.

Commission on Teacher Credentialing, California. (2003). Minutes from November 5-6, 2003 meeting, Item PERF 2: Recommended Passing Score for the California Teaching Performance Assessment (CA TPA), November 5-6, 2003. Retrieved April 5, 2009 from the web site: http://www.ctc.ca.gov/

Darling-Hammond, L. (1997). *The right to learn: A blueprint for creating schools that work.* San Francisco: Jossey-Bass Publishers.

Goodlad, J. (1994). *Educational renewal: Better teachers, better schools.* San Francisco: Jossey-Bass Publishers.

Howell, D. C. (2010). *Statistical methods for psychology*. Belmont, CA: Wadsworth Cengage.

Miles, M., & Huberman, M., (1994). *Qualitative data analysis: An expanded source book.* Thousand Oaks, CA: Sage Publications.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284-290.

Salkind, N. J. (2006). *Tests & measurement for people who (think they) hate tests & measurement.* Thousand Oaks, CA: Sage Publications.

Schultz, K. S., Riggs, M. L., & Kottke, J. L. (1999). The need for an evolving concept of validity in industrial and personnel psychology: Psychometric, legal, and emerging issues. *Current Psychology, 17*, 265-286.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 240-428.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education*. Upper Saddle River, NJ: Pearson, Merril, Prentice Hall.